

PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (<http://bmjopen.bmj.com/site/about/resources/checklist.pdf>) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Confounding adjustment methods in longitudinal observational data with a time-varying treatment: a mapping review
AUTHORS	Wijn, Stan; Rovers, Maroeska; Hannink, Gerjon

VERSION 1 – REVIEW

REVIEWER	Mohajer, Bahram Johns Hopkins University School of Medicine
REVIEW RETURNED	15-Dec-2021

GENERAL COMMENTS	<p>I read and reviewed the interesting study from Wijn et al. with pleasure. Authors have investigated an important topic using novel methods. The methodology is comprehensive and robust, and the results are in the interest of the broad group of researchers dealing with methods for reducing bias in observational studies. While extensive literature is available on methods for reducing bias associated with time-varying treatment or covariates, this study presents the current status of using these methods.</p> <p>The drawback of the study is that the authors have not considered other widely used methods for reducing bias associated with the time-varying covariates in the observational studies when defining the exposure/treatment. An excellent example of these methods is using the per-protocol approach, as suggested by Danaei et al. [Danaei, Goodarz, et al. "Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease." <i>Statistical methods in medical research</i> 22.1 (2013): 70-96.] Also, some studies have used a combination of propensity score matching (PSM), inverse probability weighting (IPW), and covariate adjustment together. However, it is unclear how these studies were handled in the literature search.</p> <p>Minor comments:</p> <ul style="list-style-type: none">• Lines 66-69, Strengths and limitations of this study: In "Although time-dependent methods like time-dependent propensity score matching, parametric g-formula, and inverse probability weighting are described in detail in the literature, adjusting at baseline in observational data is still common and was potentially inappropriate in a proportion of the papers we included in our mapping review." How is this a limitation or strength?• Lines 70-71, Strengths and limitations of this study: This sentence is a general limitation of the mapping review and not this study. "A limitation of a mapping review is the broad descriptive level at which studies are analysed. However, it does provide a general overview of the published literature."
-------------------------	---

	<ul style="list-style-type: none"> • Box 1 and Figure 1: Authors report that estimates are different according to the use of methods for covariate adjustment, but they have not assessed whether estimates are statistically significant or not. Methods like homogeneity tests can be used here. • Lines 113-114: In. "Study selection was performed by one reviewer, and issues were discussed and resolved by all authors." The selection of studies by only one reviewer can be mentioned as a limitation in the discussion.
--	---

REVIEWER	Zhang, Zhongheng Zhejiang University School of Medicine
REVIEW RETURNED	10-Jan-2022

GENERAL COMMENTS	<p>This is an interesting work to address the time-varying covariates in the effectiveness study. I have a few comments:</p> <ol style="list-style-type: none"> 1. "Confounding adjustment methods designed to deal with a time varying treatment and time varying confounding are available, but are not regularly"---this seems not fully supported by the statistics in the results section; you need to estimate the proportion of such studies. 2. The search term is problematic because many literature can describe as "marginal structural cox model" as the method employs time-varying covariates with IPW. 3. The analyses are too simple and more in-depth analysis can be done: how can different methods influence the conclusion in the original studies? for example, with time-varying covariates adjustment, the beneficial effects are less likely to be reported and more likely to report neutral effects. This point is important for methodologist 4. The specific fields of these analysis can be further explored such as oncology, cardiology, emergency care etc. 5. There can be a table to briefly describe each methods and situations when they can be used.
-------------------------	---

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

I read and reviewed the interesting study from Wijn et al. with pleasure. Authors have investigated an important topic using novel methods. The methodology is comprehensive and robust, and the results are in the interest of the broad group of researchers dealing with methods for reducing bias in observational studies. While extensive literature is available on methods for reducing bias associated with time-varying treatment or covariates, this study presents the current status of using these methods.

The drawback of the study is that the authors have not considered other widely used methods for reducing bias associated with the time-varying covariates in the observational studies when defining the exposure/treatment. An excellent example of these methods is using the per-protocol approach, as suggested by Danaei et al. [Danaei, Goodarz, et al. "Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease." *Statistical methods in medical research* 22.1 (2013): 70-96.] Also, some studies have used a combination of propensity score matching (PSM), inverse probability weighting (IPW), and covariate adjustment together. However, it is unclear how these studies were handled in the literature search.

Thank you for your comments. The reviewer points out that widely used methods for reducing bias were not considered in our paper and that this is a drawback of our study. The study of Danaei et al (2013) is quoted as an example but this study was actually included in our search. The study of Danaei et al. outlines the methodology to simulate a randomised controlled trial using observational data that is required because treatment adherence is problematic in their example. However, in the end, inverse probability weighting was used to adjust for the time-varying confounding (and was included as such in our paper).

We agree with the reviewer that it is unclear how we dealt with studies that used multiple adjustment methods or a combination of methods. Therefore, we added additional information to the method section to elaborate on how we dealt with studies using multiple adjustment methods or a combination of methods.

Analysis

Line 154: "If a study used multiple adjustment methods or a combination of methods, we included all methods, i.e., more methods than papers could be identified."

Minor comments:

- Lines 66-69, Strengths and limitations of this study: In "Although time-dependent methods like time-dependent propensity score matching, parametric g-formula, and inverse probability weighting are described in detail in the literature, adjusting at baseline in observational data is still common and was potentially inappropriate in a proportion of the papers we included in our mapping review." How is this a limitation or strength?

- Lines 70-71, Strengths and limitations of this study: This sentence is a general limitation of the mapping review and not this study. "A limitation of a mapping review is the broad descriptive level at which studies are analysed. However, it does provide a general overview of the published literature."

We agree with the reviewer that the strengths and limitations in the study summary do not properly reflect the strength or limitations of our methodology. We rewrote the article summary to reflect the strengths and limitations of the study.

Page 4: Article Summary

Strengths and limitations of this study

- We systematically mapped the literature from inception up to January 2021 for the most commonly used methods to correct for confounding in longitudinal observational data.
- This study was conducted and reported according to The PRISMA extension for Scoping Reviews (PRISMA-ScR)
- No risk of bias assessment was performed because the scope of this paper targets the statistical methods that have been used in these papers, and therefore a risk of bias assessment was not applicable.
- For some studies we were not able to identify if patients were treated at baseline or during follow-up, fortunately, this only occurred in 8% of the included papers.
- Although time-dependent methods like time-dependant propensity score matching, parametric g-formula and inverse probability weighting are described in detail in the literature, adjusting at baseline in observational data is still common and was potentially inappropriate in a proportion of the papers we included in our mapping review.
- A limitation of a mapping review is the broad descriptive level at which studies are analysed. However, it does provide a general overview of the published literature.
- Box 1 and Figure 1: Authors report that estimates are different according to the use of methods for covariate adjustment, but they have not assessed whether estimates are statistically significant or not. Methods like homogeneity tests can be used here.

The reviewer pointed out that Box 1 and Figure 1 do not include any statistical tests to determine if the results obtained with the methods differ statistically (e.g., by using a homogeneity test). We intentionally have not performed such tests as the results would be only applicable to our examples and hard to extrapolate to other studies. The aim of Box 1 and Figure 1 was to show that the methods that researchers select to correct for confounding might affect the estimated treatment effect and that it is difficult to predict the magnitude as it depends on the nature of the confounding. For example, one could find a (statistically significant) difference in effect size estimates between methods used in the first example but not in the second.

As the reviewer suggested we performed homogeneity testing and found that in the meniscectomy example (example 1) the baseline methods differed from the time-dependent methods ($p = 0.004$). However, this statistically significant difference was not present in the other example (example 2). To avoid any confusion, we decided not to include these results in the manuscript. Moreover, we do not know what the true hazard ratio in the meniscectomy example is which further limits the interpretation of a significant homogeneity test. Instead, we argue that one should start by selecting the appropriate method, and to guide the reader in selecting the appropriate method, we created a new figure (Figure 4, depicted at the bottom of this document) in which the different methods are described and when they should be used.

Results

Line 179: "We added an overview of the most commonly used methods found in our search and when they should be used."

- Lines 113-114: In. "Study selection was performed by one reviewer, and issues were discussed and resolved by all authors." The selection of studies by only one reviewer can be mentioned as a limitation in the discussion.

We agree with the reviewer that study selection by one reviewer should be included as a limitation. This has been added to the discussion.

Discussion

Line 213: Furthermore, no quality risk of bias assessment of the included studies was performed and study selection and data extraction were performed by one reviewer. Using a second reviewer throughout the entire study screening process could increase the number of relevant studies identified for use in a systematic review. (Stoll 2019) However, as we targeted the overall trends in data analysis of studies with longitudinal observational data, this would likely not affect our conclusions much.

Reviewer: 2

This is an interesting work to address the time-varying covariates in the effectiveness study. I have a few comments.

We would like to thank the reviewer for his valuable comments. We will address your comments below.

1. "Confounding adjustment methods designed to deal with a time varying treatment and time varying confounding are available, but are not regularly"---this seems not fully supported by the statistics in the results section; you need to estimate the proportion of such studies.

We agree that the conclusion "... [advanced methods] are not regularly used" currently lack the proper substantiation. From the Results section of the manuscript, we find that 45% of the papers with a

time-varying treatment used the g-methods (inverse probability weighting, parametric g-formula or g-estimation). However, we agree that this result should be described more explicitly. Therefore, we added the proportion of studies that used these advanced methods when dealing with a time-varying treatment to the abstract (conclusion), result section and the conclusion of the paper.

Abstract

Line 75: Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding are available, but were only used in 45% of the papers with a time-varying treatment. are not regularly always used.

Results

Line 172: Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding (IPW, parametric g-formula or g-estimation) were used in 45% of the papers with a time-varying treatment.

Conclusion

Line 259: Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding (IPW, parametric g-formula or g-estimation) are available, but were only used in 45% of the papers with a time-varying treatment are not regularly used and this can potentially result in biased estimates of the treatment effect.

2. The search term is problematic because many literature can describe as "marginal structural cox model" as the method employs time-varying covariates with IPW.

The reviewer is correct that adding marginal structural Cox model to the search strategy would probably help to identify studies that used time-varying covariates with IPW. Our initial search detected 227 papers that defined their methodology as such. Based on the comment of the reviewer we added "Marginal structural Cox model" to our search strategy and reran the search and found an additional 6 studies.

We have updated Supplement 1 to include the search term, screened the 6 papers full-text of which 4 were subsequently included, and have updated the manuscript accordingly.

Abstract

Line 65: "In total, 760 764 studies were included that met the eligibility criteria. PSM (165/201, 82%) and inverse probability weighting (154/502, 31% 150/498, 30%) were most common for studies with a treatment at baseline (n=201) and time-varying treatment (n=498), respectively. Of the 502 498 studies with a time-varying..."

Results:

Line 161: "Our search identified 2134 2140 articles of which eventually 760 764 met the eligibility criteria after title and abstract review, and subsequent full-text review"

Line 165: "Of all included papers, 201 (26%) had a treatment at baseline, 498 502 (66%) had a time-varying treatment and 61 (8%) papers had no clearly defined time of treatment. Of the papers with a treatment at baseline, the majority used PSM with baseline covariates (n = 165, 82%) as a method to correct for confounding. Studies that had a time-varying treatment most often used IPW (150 154 papers, 30 31%),..."

Conclusion:

Line 257: "Of the 502 498 identified studies..."

Figure 2:

Updated figure + caption:

"In total, 760 764 studies were included and categorized according to the time of treatment.

3. The analyses are too simple and more in-depth analysis can be done: how can different methods influence the conclusion in the original studies? for example, with time-varying covariates adjustment, the beneficial effects are less likely to be reported and more likely to report neutral effects. This point is important for methodologist

We agree with the reviewer that it would be interesting to analyse how different methods influence the conclusion in the original studies. However, as we do not have the source data from any of the included studies, unfortunately, we are unable to study how different methods would influence the conclusion of the original studies.

As an illustration, we have shown the influence of the different methods on the outcome in two empirical examples (found in Box 1 and Supplement 1). In these two examples, we indeed find that different adjustment methods show different effect size estimates. Choice or selection of adjustment method obviously may influence the conclusions drawn. In addition, one could imagine that authors can select a method based on effect size, which obviously may influence the conclusion of the study. Therefore, we agree that it is highly relevant to discuss these implications (i.e., potential selection of methodology) in our paper. We have added a section to the 'Implications' paragraph of the discussion in which we discuss this topic:

Discussion:

Line 249: "As we have seen in Box 1, different confounding adjustment methods can potentially influence the conclusions of a study. It depends on many (unknown) case-specific aspects and thus it can be challenging to predict how different methods can affect the conclusion of a study. A direct comparison of different methods to correct for confounding is not recommended as this could stimulate selective reporting of (positive) study results. Every analysis of longitudinal observational data should start by selecting the method best suited for the data at hand. Figure 4 provides an overview of the most commonly used methods and can assist researchers to select the most appropriate method available."

4. The specific fields of these analysis can be further explored such as oncology, cardiology, emergency care etc.

Thank you for the suggestion. We have performed additional analyses to explore our findings in specific medical fields. First, we categorized all 764 included papers by medical specialty, based on research topic. We identified 18 categories, however, 76% (n = 580) of papers fell within 5 medical specialties: Internal medicine, cardiology, urology, oncology, and geriatrics. To avoid too small categories, we focused on these 5 largest medical specialties. Please find below a plot showing the (frequency of) adjustment methods used in studies with a time-varying treatment stratified by medical specialty and a second plot showing the proportion of methods that has been used.

Figure 1: Bar chart showing the confounding adjustment methods for the 5 most common medical specialties with a time-varying treatment. TdCox, time-dependent cox regression; TdPSM, time-dependent propensity score matching.

Figure 2: Bar chart showing the proportion of confounding adjustment methods for the 5 most common medical specialties with a time-varying treatment. TdCox, time-dependent cox regression; TdPSM, time-dependent propensity score matching.

As seen from the figure, the proportions do not differ much in the use of different methods in the medical specialties. In our opinion, this stratification by medical specialty does not provide much additional information compared with our overall results. The aim of our study was to explore which confounding adjustment methods have been used in longitudinal observational data and identify potential inappropriate use of propensity score matching and these methods are currently already displayed in figure 2 and 3. We therefore decided not to include these results in the manuscript. However, if the editor and reviewer think differently, of course we are willing to put the results listed above in the manuscript or supplement.

5. There can be a table to briefly describe each methods and situations when they can be used.

Thank you for the suggestion. We have now added a figure to our manuscript that describes the different methods (and when they should be used). (the figure is also added on the last page of this letter)

Results

Line 179: "We added an overview of the most commonly used methods found in our search and when they should be used. (Figure 4)"

Line 193: "Figure 4: Common methods to correct for confounding and when they should be used."

VERSION 2 – REVIEW

REVIEWER	Mohajer, Bahram Johns Hopkins University School of Medicine
REVIEW RETURNED	10-Feb-2022
GENERAL COMMENTS	The authors addressed all the comments, and the manuscript is considerably improved. I am looking forward to reading more of their work.
REVIEWER	Zhang, Zhongheng Zhejiang University School of Medicine
REVIEW RETURNED	02-Feb-2022
GENERAL COMMENTS	no further comments